

# Universal Sequential Outlier Hypothesis Testing

Yun Li\*, Sirin Nitinawarat†, and Venugopal V. Veeravalli\*

\* Department of Electrical and Computer Engineering  
and

Coordinated Science Laboratory  
University of Illinois at Urbana-Champaign  
Urbana, IL 61801-2307, USA

Emails: {yunli2, vvv}@illinois.edu

† Qualcomm Technologies, Inc.

5775 Morehouse Drive

San Diego, CA 92121

Email: sirin.nitinawarat@gmail.com

**Abstract**—Universal outlier hypothesis testing is studied in a sequential setting. Multiple observation sequences are collected, a small subset of which are outliers. A sequence is considered an outlier if the observations in that sequence are generated by an “outlier” distribution, distinct from a common “typical” distribution governing the majority of the sequences. Apart from being distinct, the outlier and typical distributions can be arbitrarily close. The goal is to design a universal test to best discern all the outlier sequences. A universal test with the flavor of the repeated significance test is proposed and its asymptotic performance is characterized under various universal settings. The proposed test is shown to be universally consistent. For the model with identical outliers, the test is shown to be asymptotically optimal universally when the number of outliers is the largest possible and with the typical distribution being known, and its asymptotic performance otherwise is also characterized. An extension of the findings to the model with multiple distinct outliers is also discussed. In all cases, it is shown that the asymptotic performance guarantees for the proposed test when neither the outlier nor typical distribution is known converge to those when the typical distribution is known.

## I. INTRODUCTION

We consider the following inference problem of outlier hypothesis testing in a sequential setting. Among a fixed number of independent and memoryless observation sequences, it is assumed that a small subset (possibly empty) of these sequences are outliers. Specifically, most of the sequences are assumed to be distributed according to a “typical” distribution, while an outlier sequence is distributed according to an “outlier distribution,” distinct from the typical distribution. We shall be interested in a *non-parametric* setting, in which the outlier and typical distributions are not fully known and can be arbitrarily close. The goal is to design a universal test to identify all the outlier sequences using the fewest observations.

In [1], we studied universal outlier hypothesis testing in a fixed sample size setting. The main finding therein was that the generalized likelihood (GL) test is far more efficient for universal outlier hypothesis testing than for the other inference problems, such as homogeneity testing and classification [2], [3], [4]. In particular, the GL test was shown to be *universally exponentially consistent* for outlier hypothesis testing, whereas

it is impossible to achieve universally exponential consistency for homogeneity testing or classification without training data [3], [4]. We also showed that the GL test is *asymptotically optimal* in the limit of large number of sequences. Our previous paper [5] generalized the scope of these previous findings to the sequential setting, but only covered the setting with at most one outlier. In this paper, we shall focus on the extension with multiple outliers.

Sequential hypothesis testing has a rich history going back to the seminal work of Wald [6]. A majority of the results on sequential hypothesis testing have been for the case where the conditional distributions of observations under the hypotheses are completely known (see, e.g., [6], [7], [8], [9], [10], [11]). For the case where the distribution of observations is not completely specified, there have been a number of results for composite hypothesis testing with parametric families of distributions [12], [13]. As elucidated by Wald [6], there are two general approaches for constructing sequential tests for such parametric settings, one based on a weighted (or mixture) likelihood function for each hypothesis (see, e.g., [12]), and the other based on a maximum (generalized) likelihood function for each hypothesis (see, e.g., [13]). There have also been a limited number of papers on non-parametric approaches to sequential hypothesis testing where the functional form of the distribution is unknown, but it is known, for example, that the conditional distribution under the various hypotheses are rigid translations of each other (see, e.g., [14]). Sequential outlier hypothesis testing is closely related to the so called *slippage problem* studied in the sequential setting (see, e.g., [15]). In the slippage problem,  $N$  populations are identically distributed except possibly for one. The goal is to decide whether or not one of the populations has “slipped”, if so, which one. However, such prior work on the slippage problem concerned the situation where the typical and “slipped” distributions are tightly coupled, for example, when they are mean-shifted versions of each other. In universal sequential outlier hypothesis testing, we have no information regarding the outlier and typical distributions. In particular, the typical and outlier distributions can be arbitrarily distributed and they

can be arbitrarily close to each other. In addition, we have no training data to learn these distributions before the test is performed. To the best of our knowledge, there has been no prior work on sequential outlier hypothesis testing in such a fully non-parametric setting that we study in this paper. A key assumption that we make is that each instantaneous observation takes value in a finite and known set. Under this assumption, we shall construct an efficient universal test that will be proven to be universally exponentially consistent, and to be sometimes optimal universally or in the limit of large number of sequences. The proposed universal test has the flavor of the repeated significance test [16], [17], wherein the test stops when the generalized likelihood for the most likely hypothesis is larger by a time-dependent threshold than those for all the competing hypotheses, if that happens before a certain time.

In Section II, we start with definitions of relevant distances between pairs of distributions, key to our results. Sections III, IV concern the models with identical and distinctly distributed outliers, respectively. Performance of our proposed tests is evaluated on real data relevant to spam detection applications in Section V. Due to space limitations, proofs of our results are omitted.

## II. PRELIMINARIES

Throughout the paper, random variables (rvs) are denoted by capital letters, and their realizations are denoted by the corresponding lower-case letters. All rvs are assumed to take values in *finite* sets, and all logarithms are the natural one.

Our results will be stated in terms of certain distance metrics between a pair of distributions  $p, q$  on  $\mathcal{Y}$ : the *Bhattacharyya distance* and the *relative entropy*, denoted by  $B(p, q)$  and  $D(p||q)$ , respectively, and defined as (see, e.g., [18])

$$B(p, q) \triangleq -\log \left( \sum_{y \in \mathcal{Y}} p(y)^{\frac{1}{2}} q(y)^{\frac{1}{2}} \right),$$

and

$$D(p||q) \triangleq \sum_{y \in \mathcal{Y}} p(y) \log \frac{p(y)}{q(y)},$$

respectively.

## III. MODEL WITH IDENTICAL OUTLIERS

Consider  $M \geq 3$  independent sequences, each of which consists of independent and identically distributed (i.i.d.) observations. Denote the  $k$ -th observation of the  $i$ -th sequence by  $Y_k^{(i)} \in \mathcal{Y}$ . We assume that there are up to  $K > 2$  outliers among the  $M$  sequences with  $K < \frac{M}{2}$ , and that each of the outliers are identically distributed (i.i.d.) according to  $\mu \in \mathcal{P}(\mathcal{Y})$ , whereas all the other sequences are distributed according to the typical distribution  $\pi \in \mathcal{P}(\mathcal{Y})$ . Under the hypothesis with no outlier, namely, the *null hypothesis*, all sequences are commonly distributed according to the typical distribution. *Nothing is known about  $\mu$  and  $\pi$  except that  $\mu \neq \pi$ , and that each of them has a full support.* The

assumption of  $\mu, \pi$  having full supports rules out trivial cases where it is easier to identify the outlier sequences.

It was shown in [1] that in the fixed sample size setting, this assumption of the outliers being identically distributed is essential for the existence of a test that is universally exponentially consistent (under all the non-null hypotheses) when the number of outliers is not completely specified (anything from 0 to  $K$ ). In the next Section IV, we shall look at the extension with possibly distinctly distributed outliers but with their total number being known.

When there are some outliers, with the set of all outliers denoted by  $S$ ,  $0 < |S| < \frac{M}{2}$ , the joint distribution of the first  $n$  observations is given by

$$\begin{aligned} p_S(\mathbf{y}^n) &= p_S(\mathbf{y}_1, \dots, \mathbf{y}_n) \\ &= \prod_{k=1}^n \left\{ \prod_{i \in S} \mu(y_k^{(i)}) \prod_{j \notin S} \pi(y_k^{(j)}) \right\}. \end{aligned} \quad (1)$$

Under the null hypothesis with no outlier, the joint distribution of the observations is given as

$$p_0(\mathbf{y}^n) = p_\emptyset(\mathbf{y}^n) = \prod_{k=1}^n \prod_{i=1}^M \pi(y_k^{(i)}).$$

A sequential test for the outlier consists of a stopping rule and a final decision rule. The stopping rule defines a random (Markov) time, denoted by  $N$ , which is the number of observations taken until a final decision is made. At the stopping time  $N = n$ , a decision is made based on a decision rule  $\delta : \mathcal{Y}^{Mn} \rightarrow \mathcal{S}$ , where  $\mathcal{S}$  denotes the set of all subsets of  $\{1, \dots, M\}$  of size at most  $K$ . The overall goal of sequential testing is to achieve a certain level of accuracy for the final decision using the fewest number of observations on average.

We consider the sequential outlier hypothesis testing problem in two settings: the setting where only  $\pi$  is known, and the completely universal setting where neither  $\mu$  nor  $\pi$  is known. Consequently, a universal test is not allowed to be a function of  $\mu$ , and of  $\mu$  or  $\pi$ , in the respective settings.

The accuracy of a sequential test is gauged using the maximal error probability  $P_{\max}$ , which is a function of both the test and  $(\mu, \pi)$  and is defined as

$$P_{\max} \triangleq \max_{S \in \mathcal{S}} \mathbb{P}_S \left\{ \delta(\mathbf{Y}^N) \neq S \right\}, \quad (2)$$

where  $\mathbb{P}_S$ ,  $S \in \mathcal{S}$ , denotes the probability distribution for the hypothesis with  $S$  as the subset of all outliers. We say a sequence of tests is *universally consistent* if the maximal error probability converges to zero for any  $\mu, \pi, \mu \neq \pi$ . Further, we say it is *universally exponentially consistent* if the exponent for the maximal error probability with respect to the expected stopping time under each hypothesis is strictly positive, i.e., there exists  $\alpha_S > 0$  such that

$$\mathbb{E}_S[N] \leq \frac{-\log P_{\max}}{\alpha_S} (1 + o(1)) \quad (3)$$

for any  $\mu, \pi, \mu \neq \pi$  as  $P_{\max} \rightarrow 0$ .

We first consider the setting where both the typical and outlier distributions are known. In this non-universal setting, the Multihypothesis Sequential Probability Ratio Test (MSPRT) was shown to be asymptotically optimal in the regime with vanishing error probability [10]. For a given threshold  $T > 0$  and with  $\hat{S}(\mathbf{y}^n) \triangleq \underset{S \in \mathcal{S}}{\operatorname{argmax}} p_S(\mathbf{y}^n)$ , denoting the instantaneous maximum likelihood (ML) estimate of the hypothesis at time  $n$ , the stopping time  $N^*$  and final decision  $\delta^*$  of the MSPRT are defined as follows.

$$N^* = \underset{n \geq 1}{\operatorname{argmin}} \left[ \frac{p_{\hat{S}}(\mathbf{Y}^n)}{\max_{S \neq \hat{S}} p_S(\mathbf{Y}^n)} > T \right], \quad (4)$$

$$\delta^* = \hat{S}(\mathbf{Y}^{N^*}). \quad (5)$$

*Proposition 1:* As the threshold  $T$  in (4) approaches infinity, the MSPRT in (4), (5) satisfies  $P_{\max} = O(\frac{1}{T})$ . In addition, it yields that

$$\mathbb{E}_S[N^*] = \begin{cases} \frac{-\log P_{\max}}{D(\mu\|\pi)}(1+o(1)), & |S| = K; \\ \frac{-\log P_{\max}}{\min\{D(\mu\|\pi), D(\pi\|\mu)\}}(1+o(1)), & 1 \leq |S| < K; \\ \frac{-\log P_{\max}}{D(\pi\|\mu)}(1+o(1)), & S = \emptyset. \end{cases}$$

Furthermore, the MSPRT is asymptotically optimal. In particular, for any sequence of tests  $(N, \delta)$  with vanishing maximal error probability, it holds (simultaneously) that

$$\mathbb{E}_S[N] \geq \begin{cases} \frac{-\log P_{\max}}{D(\mu\|\pi)}(1+o(1)) & |S| = K; \\ \frac{-\log P_{\max}}{\min\{D(\mu\|\pi), D(\pi\|\mu)\}}(1+o(1)) & 1 \leq |S| < K; \\ \frac{-\log P_{\max}}{D(\pi\|\mu)}(1+o(1)) & S = \emptyset. \end{cases}$$

Now we consider the universal settings when the outlier distribution is unknown, and when neither the outlier nor typical distribution is known. In the fixed sample size setting, it was shown in [1] that a universally exponentially consistent test cannot exist. Correspondingly, we proposed a test therein that fulfilled a lesser objective of attaining universally exponential consistency under all the non-null hypotheses, while satisfying *only* universal consistency under the null hypothesis. We now describe a universal sequential test fulfilling a similar objective.

1) *Proposed Universal Test:* For each  $i = 1, \dots, M$ , denote the empirical distribution of  $\mathbf{y}^{(i)}$  by  $\gamma_i$ . When only  $\pi$  is known, we compute the generalized likelihood of  $\mathbf{y}^n$  under each non-null hypothesis corresponding to a non-empty subset  $S \subset \{1, \dots, M\}$  by replacing the unknown  $\mu$  in (1) with its ML estimate  $\hat{\mu}_S \triangleq \frac{\sum_{i \in S} \gamma_i}{|S|}$ , as

$$\hat{p}_S^{\text{typ}}(\mathbf{y}^n) = \prod_{k=1}^n \left\{ \prod_{i \in S} \hat{\mu}_S(y_k^{(i)}) \prod_{j \notin S} \pi(y_k^{(j)}) \right\}. \quad (6)$$

Similarly, when neither  $\pi$  nor  $\mu$  is known, we compute the generalized likelihood of  $\mathbf{y}^n$  under each non-null hypothesis corresponding to a non-empty  $S \in \mathcal{S}$  by replacing the un-

known  $\mu$  and  $\pi$  in (1) with their ML estimates  $\hat{\mu}_S \triangleq \frac{\sum_{i \in S} \gamma_i}{|S|}$ , and  $\hat{\pi}_S \triangleq \frac{\sum_{j \notin S} \gamma_j}{M-|S|}$ , respectively, as

$$\hat{p}_S^{\text{univ}}(\mathbf{y}^n) = \prod_{k=1}^n \left\{ \prod_{i \in S} \hat{\mu}_S(y_k^{(i)}) \prod_{j \notin S} \hat{\pi}_S(y_k^{(j)}) \right\}. \quad (7)$$

When only  $\pi$  is known and with  $\hat{S}(\mathbf{Y}^N) \triangleq$

$$\underset{\substack{S \in \mathcal{S} \\ S \neq \emptyset}}{\operatorname{argmax}} \hat{p}_S^{\text{typ}}(\mathbf{y}^n) = \underset{\substack{S \in \mathcal{S} \\ S \neq \emptyset}}{\operatorname{argmin}} \left[ \sum_{i \in S} D(\gamma_i \| \frac{\sum_{k \in S} \gamma_k}{|S|}) + \sum_{j \notin S} D(\gamma_j \| \pi) \right]$$

denoting the instantaneous estimate of the non-null hypothesis (using the generalized likelihood) at time  $n$ , our proposed universal test can be described by the following stopping and final decision rules

$$N^* = \min(\tilde{N}, \lfloor f(T) \rfloor), \quad (8)$$

$$\delta^* = \begin{cases} \hat{S}(\mathbf{Y}^{N^*}) & \text{if } \tilde{N} \leq f(T); \\ 0 & \text{if } \tilde{N} > f(T), \end{cases} \quad (9)$$

where  $f(T)$  is any function growing at least as fast as  $T \log T$ , and

$$\begin{aligned} \tilde{N} \triangleq \underset{n \geq 1}{\operatorname{argmin}} & \left[ \min_{\substack{S' \neq \tilde{S} \\ S' \neq \emptyset}} n \left[ \sum_{i \in S'} D(\gamma_i \| \frac{\sum_{k \in S'} \gamma_k}{|S'|}) + \sum_{j \notin S'} D(\gamma_j \| \pi) \right. \right. \\ & \left. \left. - \sum_{i \in \tilde{S}} D(\gamma_i \| \frac{\sum_{k \in \tilde{S}} \gamma_k}{|\tilde{S}|}) - \sum_{j \notin \tilde{S}} D(\gamma_j \| \pi) \right] \right. \\ & \left. > \log T + (M+1)|\mathcal{Y}| \log(n+1) \right], \quad (10) \end{aligned}$$

Similarly, when neither  $\mu$  nor  $\pi$  is known, the test can be written as in (8), (9) but with  $\hat{S}(\mathbf{Y}^N) \triangleq \underset{S \in \mathcal{S}, S \neq \emptyset}{\operatorname{argmax}} \hat{p}_S^{\text{univ}}(\mathbf{y}^n) = \underset{S \in \mathcal{S}, S \neq \emptyset}{\operatorname{argmin}} \left[ \sum_{i \in S} D(\gamma_i \| \frac{\sum_{k \in S} \gamma_k}{|S|}) + \sum_{j \notin S} D(\gamma_j \| \frac{\sum_{k \notin S} \gamma_k}{M-|S|}) \right]$ , and

$$\begin{aligned} \tilde{N} \triangleq \underset{n \geq 1}{\operatorname{argmin}} & \left[ \min_{\substack{S' \neq \tilde{S} \\ S' \neq \emptyset}} n \left[ \sum_{i \in S'} D(\gamma_i \| \frac{\sum_{k \in S'} \gamma_k}{|S'|}) \right. \right. \\ & \left. \left. + \sum_{j \notin S'} D(\gamma_j \| \frac{\sum_{k \notin S'} \gamma_k}{M-|S'|}) \right. \right. \\ & \left. \left. - \sum_{i \in \tilde{S}} D(\gamma_i \| \frac{\sum_{k \in \tilde{S}} \gamma_k}{|\tilde{S}|}) \right. \right. \\ & \left. \left. - \sum_{j \notin \tilde{S}} D(\gamma_j \| \frac{\sum_{k \notin \tilde{S}} \gamma_k}{M-|\tilde{S}|}) \right] \right. \\ & \left. > \log T + (M+1)|\mathcal{Y}| \log(n+1) \right]. \quad (11) \end{aligned}$$

2) *Performance of Proposed Test:*

*Theorem 2:* When only  $\pi$  is known, the test in (8), (9), (10) is universally consistent, and yields for every  $T$  that  $P_{\max} \leq \frac{C}{T}$ , where  $C$  is a constant that depends only on  $M, K, \mu, \pi$ , but

not on  $T$ . In addition, it satisfies for each non-null hypothesis  $S \in \mathcal{S}, S \neq \emptyset$ , that as  $T \rightarrow \infty$ ,

$$\begin{aligned} \mathbb{E}_S[N^*] &\leq \frac{\log T}{\alpha_S}(1 + o(1)) \\ &\leq \begin{cases} \frac{-\log P_{\max}}{D(\mu\|\pi)}(1 + o(1)), & |S| = K; \\ \frac{-\log P_{\max}}{\min(D(\mu\|\pi), \eta_S(\mu\|\pi))}(1 + o(1)), & 1 \leq |S| < K, \end{cases} \end{aligned} \quad (12)$$

where

$$\begin{aligned} \alpha_S \triangleq \min_{\substack{S' \neq S \\ S' \neq \emptyset}} &\left[ |S \cap S'| D\left(\mu \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|}\right.\right) \right. \\ &+ |S \setminus S'| D(\mu\|\pi) \\ &\left. + |S' \setminus S| D\left(\pi \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|}\right.\right) \right] > 0. \end{aligned}$$

and

$$\eta_S(\mu\|\pi) \triangleq \min_{p \in \mathcal{P}(\mathcal{Y})} |S| D(\mu\|p) + D(\pi\|p). \quad (13)$$

*Theorem 3:* When neither  $\mu$  nor  $\pi$  is known, the universal test in (8), (9), (11) is universally consistent, and yields for every  $T$  that  $P_{\max} \leq \frac{C}{T}$ , where  $C$  is a constant that depends on  $M, K, \mu, \pi$ , but not on  $T$ . In addition, for each non-null hypothesis  $S \in \mathcal{S}, S \neq \emptyset$ , as  $T \rightarrow \infty$ ,

$$\begin{aligned} \mathbb{E}_S[N^*] &\leq \frac{\log T}{\bar{\alpha}_S}(1 + o(1)) \\ &\leq \begin{cases} \frac{-\log P_{\max}}{\bar{\eta}(\mu\|\pi)}(1 + o(1)), & |S| = K; \\ \frac{-\log P_{\max}}{\min(\bar{\eta}(\mu\|\pi), \eta_S(\mu\|\pi))}(1 + o(1)), & 1 \leq |S| < K, \end{cases} \end{aligned} \quad (15)$$

where

$$\begin{aligned} \bar{\alpha}_S \triangleq \min_{\substack{S' \neq S \\ S' \neq \emptyset}} &\left[ |S \cap S'| D\left(\mu \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|}\right.\right) \right. \\ &+ |S \setminus S'| D\left(\mu \left\| \frac{|S \setminus S'| \mu + |S' \cap S^c| \pi}{M - |S'|}\right.\right) \\ &+ |S' \setminus S| D\left(\pi \left\| \frac{|S \cap S'| \mu + |S' \setminus S| \pi}{|S'|}\right.\right) \\ &\left. + |S^c \cap S'^c| D\left(\pi \left\| \frac{|S \setminus S'| \mu + |S' \cap S^c| \pi}{M - |S'|}\right.\right) \right] > 0, \end{aligned}$$

and

$$\bar{\eta}_S(\mu\|\pi) \triangleq \min_{p \in \mathcal{P}(\mathcal{Y})} D(\mu\|p) + (M - K - |S|) D(\pi\|p), \quad (16)$$

and  $\eta_S(\mu\|\pi)$  is as in (13).

*Remark 1:* It follows from (16) that as  $M \rightarrow \infty$ ,

$$\bar{\eta}_S(\mu, \pi) \rightarrow D(\mu\|\pi), \quad (17)$$

i.e., the asymptotic performance guarantee for the test in (8), (9), (11) when neither  $\mu$  nor  $\pi$  (cf. (15)) are known converges to that for the test in (8), (9), (10) when  $\pi$  is known (cf. (12)) as  $M \rightarrow \infty$ .

#### IV. MODEL WITH DISTINCT OUTLIERS

It was shown in [1] that when the outliers can be arbitrarily distinctly distributed, the assumption of the number of outliers being known is essential for the existence of a universally exponentially consistent test. We now describe this extension with distinctly distributed outliers but with their number being known in the sequential setting.

In particular, for an  $S \subset \{1, \dots, M\}$ ,  $|S| = K$ , denoting the set of  $K$  outliers, the joint distribution of all observations under the hypothesis with the outlier subset being  $S$  is

$$\begin{aligned} p_S(\mathbf{y}^n) &= p_S(\mathbf{y}_1, \dots, \mathbf{y}_n) \\ &= \prod_{k=1}^n \left\{ \prod_{i \in S} \mu_i(y_k^{(i)}) \prod_{j \notin S} \pi(y_k^{(j)}) \right\}, \end{aligned} \quad (18)$$

where each  $i$ -th outlier,  $i \in S$ , is distributed as  $\mu_i$ , which can be arbitrarily distinct from one another as long as each  $\mu_i \neq \pi$ . At the stopping time  $N = n$ , the test for the outliers is done based on a rule  $\delta : \mathcal{Y}^{Mn} \rightarrow \mathcal{S}_K$ , where  $\mathcal{S}_K$  will now denote the set of all subsets of  $\{1, \dots, M\}$  of size *exactly*  $K$ . Notice that unlike in the previous sections, the current model does not include the null hypothesis with no outlier. The maximal error probability is defined as previously in (2) but with the maximum being over  $\mathcal{S}_K$  instead.

##### A. Proposed Universal Test

When only  $\pi$  is known, we can compute the corresponding generalized likelihood of  $\mathbf{y}^n$  under each hypothesis  $S \in \mathcal{S}_K$  by replacing the unknown  $\mu_i$ ,  $i \in S$ , in (18) with its ML estimate  $\hat{\mu}_S^i \triangleq \gamma_i$ . In particular, with  $\hat{S}(\mathbf{Y}^n) = \operatorname{argmin}_{S \in \mathcal{S}_K} \sum_{j \notin S} D(\gamma_j\|\pi)$  denoting the instantaneous estimate of the hypothesis (using the generalized likelihood) at time  $n$ , our proposed universal test can be described by the following stopping and final decision rules:

$$\begin{aligned} N^* &= \operatorname{argmin}_{n \geq 1} \left[ \min_{\substack{S' \neq \hat{S} \\ S' \in \mathcal{S}_K}} n \left[ \sum_{j \notin S'} D(\gamma_j\|\pi) - \sum_{j \notin \hat{S}} D(\gamma_j\|\pi) \right] \right. \\ &\quad \left. > \log T + (M + 1)|\mathcal{Y}| \log(n + 1) \right]; \end{aligned} \quad (19)$$

$$\delta^* = \hat{S}(\mathbf{Y}^{N^*}). \quad (20)$$

Similarly, when neither  $\mu$  nor  $\pi$  is known, the test can be written as

$$\begin{aligned} N^* &= \operatorname{argmin}_{n \geq 1} \left[ \min_{\substack{S' \neq \hat{S} \\ S' \in \mathcal{S}_K}} n \left[ \sum_{j \notin S'} D\left(\gamma_j \left\| \frac{\sum_{k \notin S'} \gamma_k}{M - |S'|} \right.\right) \right. \right. \\ &\quad \left. \left. - \sum_{j \notin \hat{S}} D\left(\gamma_j \left\| \frac{\sum_{k \notin \hat{S}} \gamma_k}{M - |\hat{S}|} \right.\right) \right] \right. \\ &\quad \left. > \log T + (M + 1)|\mathcal{Y}| \log(n + 1) \right]; \end{aligned} \quad (21)$$

$$\delta^* = \hat{S}(\mathbf{Y}^{N^*}), \quad (22)$$



but with  $\hat{S}(Y^n) = \underset{S \in \mathcal{S}}{\operatorname{argmin}} \sum_{j \notin S} D\left(\gamma_j \parallel \frac{\sum_{k \notin S} \gamma_k}{M - |S|}\right)$ . Note that since the null hypothesis is not present in this case, there is no need to truncate the stopping time by a predefined horizon as in (8).

### B. Performance of the Proposed Tests

**Theorem 4:** With the number of distinct outliers  $K$  being known and when only  $\pi$  is known, the test in (19), (20) is universally exponentially consistent, and yields for every  $T$  that  $P_{\max} \leq \frac{C}{T}$ , where  $C$  is a constant that depends only on  $M, K, \mu, \pi$ , but not on  $T$ . In addition, for each hypothesis  $S \in \mathcal{S}_K$ , as  $T \rightarrow \infty$ ,

$$\mathbb{E}_S[N^*] \leq \frac{-\log P_{\max}}{\min_{i \in S} D(\mu_i \parallel \pi)} (1 + o(1)). \quad (23)$$

**Theorem 5:** With the number of distinct outliers  $K$  being known, but neither  $\mu$  nor  $\pi$  being known, the test in (21), (22) is universally exponentially consistent, and yields for every  $T$  that  $P_{\max} \leq \frac{C}{T}$ , where  $C$  is a constant that depends only on  $M, K, \mu, \pi$ , but not on  $T$ . In addition, for each hypothesis  $S \in \mathcal{S}_K$ , as  $T \rightarrow \infty$ ,

$$\mathbb{E}_S[N^*] \leq \frac{-\log P_{\max}(1 + o(1))}{\min_{i \in S} \min_p (D(\mu_i \parallel p) + (M - 2K)D(\pi \parallel p))}. \quad (24)$$

**Remark 2:** As  $M \rightarrow \infty$ , the inner minimum in the denominator in (24) is attained at  $p^* = \pi$  and, hence, the coefficient therein converges to  $\min_{i \in S} D(\mu_i \parallel p)$ , which is the asymptotic performance of the universal test in (19), (20) when  $\pi$  is known (cf. (23)).

## V. APPLICATION TO SPAM DETECTION

We evaluate the performance of the proposed universal tests on real data set relevant to spam detection applications. The labeled data set (spam or non-spam) contains information from 4610 emails addressed to an employee at Hewlett-Packard and is publicly available [19]. In particular, the data set consists of relative frequencies of a set of 48 words and 6 punctuation marks. Out of 4601 emails, there are 1813 spams.

We design an experiment for the case with  $M = 5$  sequences, and with at most two identical outliers for the total number hypotheses of 16. Instead of looking at the frequencies of all words and punctuation marks available in the data set, we just pick out the frequency of the word “HP,” and quantize it into 5 levels (in the original data set, the frequencies take continuous values in the interval  $[0, 100]$ ). The  $f(T)$  in (8), (9) is selected to be  $T^5$ . The values of  $\frac{-\log P_{\max}}{\mathbb{E}_S[N^*]}$ , for  $|S| = 1$ , and  $|S| = 2$ , achievable by the universal test in (8), (9), (11) (without knowledge of either  $\mu$  or  $\pi$ ) is listed as a function of  $T$  in Table I. The asymptote of the expected stopping time (relative to the exponent for the error probability) under a hypothesis with  $|S| = 2$  is lower than that under a hypothesis with  $|S| = 1$ , which agrees with the results in (14) and (15).

TABLE I

	$T = 3.98$	$T = 4$	$T = 4.05$	$T = 4.1$
$\frac{-\log P_{\max}}{\mathbb{E}_{\{1\}}[N^*]}$	0.0012	0.0017	0.0039	0.0069
$\frac{-\log P_{\max}}{\mathbb{E}_{\{1,2\}}[N^*]}$	0.0017	0.0025	0.0057	0.01

## ACKNOWLEDGMENT

This research was partially supported by the Air Force Office of Scientific Research (AFOSR) under Grant FA9550-10-1-0458 through the University of Illinois at Urbana-Champaign, and by the National Science Foundation under Grant NSF CCF 11-11342.

## REFERENCES

- [1] Y. Li, S. Nitinawarat, and V. V. Veeravalli, “Universal outlier hypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 60, pp. 4066 – 4082, 2014.
- [2] K. Pearson, “On the probability that two independent distributions of frequency are really samples from the same population,” *Biometrika*, vol. 8, pp. 250–254, 1911.
- [3] J. Ziv, “On classification with empirically observed statistics and universal data compression,” *IEEE Trans. Inf. Theory*, vol. 34, pp. 278–286, 1988.
- [4] M. Gutman, “Asymptotically optimal classification for multiple tests with empirically observed statistics,” *IEEE Trans. Inf. Theory*, vol. 35, pp. 401–408, 1989.
- [5] Y. Li, S. Nitinawarat and V. V. Veeravalli, “Universal sequential outlier hypothesis testing,” in *Proc. IEEE Int. Symp. Inf. Theory*, Jun. 29-Jul. 4 2014, pp. 3205–3209.
- [6] A. Wald, “Sequential tests of statistical hypotheses,” *Ann. Math. Statist.*, vol. 16, pp. 117–186, 1945.
- [7] A. Wald and J. Wolfowitz, “Optimum character of the sequential probability ratio test,” *Ann. Math. Statist.*, vol. 19, pp. 326–339, 1948.
- [8] C. W. Baum and V. V. Veeravalli, “A sequential procedure for multihypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 40, pp. 1994–2007, Nov. 1994.
- [9] V. V. Veeravalli and C. W. Baum, “Asymptotic efficiency of a sequential multihypothesis testing,” *IEEE Trans. Inf. Theory*, vol. 41, pp. 1994–1997, Nov. 1995.
- [10] V. P. Dragalin, A. G. Tartakovsky, V. V. Veeravalli, “Multihypothesis sequential probability ratio tests—part I: Asymptotic optimality,” *IEEE Trans. Inf. Theory*, vol. 45, pp. 2448–2461, Nov. 1999.
- [11] —, “Multihypothesis sequential probability ratio tests—part II: Accurate asymptotic expansions for the expected sample size,” *IEEE Trans. Inf. Theory*, vol. 46, pp. 1136–1383, Jul. 2000.
- [12] S. Zacks, *Theory of Statistical Inference (Probability and Mathematical Statistics)*. John Wiley and Sons, Inc., 1971.
- [13] T. L. Lai, “Nearly optimal sequential tests of composite hypotheses,” *Ann. Statist.*, vol. 16, pp. 856–886, 1988.
- [14] F. Mosteller, “A  $k$ -sample slippage test for an extreme population,” *Ann. Math. Statist.*, vol. 19, pp. 58–65, 1948.
- [15] T. S. Ferguson, *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, 1967.
- [16] M. Woodroffe, *Nonlinear Renewal Theory in Sequential Analysis*, ser. CBMS-NSF regional conference series in applied mathematics. SIAM, 1982.
- [17] D. Siegmund, *Sequential Analysis: Tests and Confidence Intervals*, ser. Springer series in statistics. Springer-Verlag, 1985.
- [18] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: John Wiley and Sons, Inc., 2006.
- [19] T. Hastie, R. Tibshirani, and J. Friedman, *Elements in Statistical Learning: Data Mining, Inference, and Prediction*. Springer, <http://statweb.stanford.edu/tibs/ElemStatLearn/>, 2009.